A TOOL FOR REENGINEERING LARGE DATABASES

Mark C. Wagner (mwagner@llnl.gov), Shannon Waller, T. Mimi Yeh, and Thomas R., Slezak.
Human Genome Center, Biology and Biotechnology Research Division, Lawrence Livermore National Laboratory, Livermore CA 94550.

The Human Genome Center at Lawrence Livermore National Laboratory has been developing and using a relational database over the last 6 years. This database was designed to support our immediate need to hold only our data for human chromosome 19. We have accomplished our physical mapping task for chromosome 19 and are now poised to map and sequence other areas of human and non-human genomes. Since our current database is human chromosome 19 specific, a new database schema was required to accomplish these goals (see the poster by Tom Slezak, et. al. for details). This schema uses a large degree of abstraction and different ways of organizing the data in the original database, reducing the table count from 200 to approximately 150.

Reengineering a large database and populating it with the contents of the original is a complex task, certain to require many iterations. We needed a tool specific to this task that would let us specify all details of the new database in special "meta tables", plus routines which could automatically generate all the table, rule, trigger, index, user, and permission SQL. A byproduct of this tool would be on-line and current documentation, as well as table-specific backup dump procedures. We have also had to devise a method to "translate" our data from our existing database into our new schema. This is done at column-level granularity: for each column in the new database, the SQL necessary to fill that column is specified. In this fashion, if a column is moved or reordered, a simple automated routine can readily extract the data for uploading into the new table. Changes to the schema are effected in the meta tables. The process of generating and populating the new database from the old is completely automated.

We used PERL and HTML to create a WWW graphical interface to develop this translation tool. This permitted us to develop one interface for use across multiple platforms with a minimum of effort.

It should be noted that this tool was not designed for implementing a database from scratch, although it could be used in that application. The strength of this tool lies in its ability to move data from one database schema to another with a minimum amount of human intervention.

(This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.)